Bacterial riboproteogenomics: the era of N-terminal proteoform existence revealed

Daria Fijalkowska[1Ŧ], Igor Fijalkowski[1Ŧ], Patrick Willems[1] and Petra Van Damme[1*]

[1] Department of Biochemistry and Microbiology, Ghent University, K. L. Ledeganckstraat 35, B-9000 Ghent, Belgium.

* Correspondence: Petra.VanDamme@ugent.be, Phone: +32 92649279.

Ŧ Authors contributed equally to this work

Keywords

(alternative) translation initiation, bacterial genome annotation, riboproteogenomics, N-terminal proteoforms

Abstract

With the rapid increase in the number of sequenced prokaryotic genomes, relying on automated gene annotation became a necessity. Multiple lines of evidence, however, suggest that current bacterial genome annotations may contain inconsistencies and are incomplete, even for so-called well-annotated genomes. We here discuss underexplored sources of protein diversity and new methodologies for high-throughput genome re-annotation. The expression of multiple molecular forms of proteins (proteoforms) from a single gene, particularly driven by alternative translation initiation, is gaining interest as a prominent contributor to bacterial protein diversity. In consequence, riboproteogenomic pipelines were proposed to comprehensively capture proteoform expression in prokaryotes by the complementary use of (positional) proteomics and the direct readout of translated genomic regions using ribosome profiling. To complement these discoveries, tailored strategies are required for the functional characterization of newly discovered bacterial proteoforms.

Challenges in bacterial genome annotation

Whole genome sequencing efforts in microbiology are mounting at an unprecedented pace. With over 200.000 prokaryotic genomes currently available (NCBI, July 2019), genome sequence analysis and gene annotation becomes a challenging task requiring automated gene prediction. As of 2018, curation efforts within NCBI have almost exclusively shifted to fine-tuning annotation rules for

automatic pipelines rather that manual curation of existing genome annotations (Haft et al. 2018). Despite efforts to implement unified analysis strategies for newly sequenced organisms or updated assemblies (Haft et al. 2018; Tatusova et al. 2016), for historical reasons, significant differences in genome annotation pipelines applied for some classic model organisms caused propagation of gene annotation inconsistencies between related species and strains (Blattner et al. 1997; Jarvik et al. 2010; Kroger et al. 2012; McClelland et al. 2001; Poptsova and Gogarten 2010; Tatusova et al. 2016). The pipelines, such as NCBI's PGAP (Prokaryotic Genome Annotation Pipeline), include collections of Hidden Markov models (HMMs) trained to detect specific families of genes sharing common structural or sequence properties (Finn et al. 2016; Haft et al. 2013). Albeit useful, such methods often fail to detect novel protein families as evident from a recent effort identifying genes encoding surface lipoprotein assembly modulators in 638 bacterial species, many of which were not previously reported to possess genes of this family (Hooda et al. 2017). Moreover, still not all genomes submitted into repositories are annotated using state-of-the-art pipelines, such as PGAP, and standardization of procedures remains difficult given the biological diversity within Prokaryota (Tripp et al. 2015). With the plethora of annotation tools available to researchers, each intrinsically burdened with specific biases, major challenges still persist in functional genome annotation as evident from the poor (~70%) agreement in predictions made by popular tools such as Genemark, Glimmer and Prodigal for a number of genomes (Tripp et al. 2015). These discrepancies persist despite continuous development of prediction tools, including the novel modelling of leaderless and atypical genes in GenemarkS2 (Lomsadze et al. 2018). The historical scarcity of high quality, experimentally verified gene sets, such as EcoGene verified set of sequenced protein starts, hindered the evaluation of the performance of gene prediction tools (Hyatt et al. 2010; Rudd 2000). Therefore, continuous efforts to update gene and ORF annotations are of utmost importance (Bland et al. 2014; Haft et al. 2018; Poptsova and Gogarten 2010). In this context, recent computational methods were developed that rely on experimental translation data to predict protein-coding regions in bacteria (Clauwaert et al. 2019; Giess et al. 2017; Ndah et al. 2017).

In the absence of experimental data, structural and functional gene annotation relies on *ab initio* gene prediction and homology to well-characterized reference genomes. Although prediction based on sequence similarity is generally a valid strategy, it may propagate incorrect gene structures and functions across phylogeny (Poptsova and Gogarten 2010; Promponas et al. 2015). In addition, the use of sequence features, like GC-content, to select model parameters skews gene delineation leading to the observed underestimation of the proteome complexity (Kelley et al. 2012). Consequently, performance of prediction algorithms varies widely for genomes displaying different GC content, both in overall gene prediction accuracy as well as translation start site detection (Tripp

et al. 2015). A large number of spurious, typically short, open reading frames (ORFs) predicted by such tools often does not represent real gene products (Marcellin et al. 2013). In order to reduce erroneous predictions, arbitrary cut-offs have to be used when considering the minimal lengths of putative genes. In case of Glimmer, such recommended cut-off is 120 nucleotides, preventing the tool from detecting some well characterized small genes (e.g. the 39 nucleotide long ORF encoding PatS) (Kelley et al. 2012). In this context it is noteworthy that a growing body of evidence indicates that this particular class of genes suffers significantly from underannotation despite their proven biological relevance (Impens et al. 2017; Lloyd et al. 2017; Miravet-Verde et al. 2019; Sberro et al. 2019). More specifically, in a recent effort aimed at reannotating the genome of *Trypanosoma brucei*, the median size of 225 newly identified translation products was only 81 aa (Parsons et al. 2015). All the shortcomings of currently used automated annotation tools highlight the importance of context-specific experimental validation and biological evidence-based genome reannotation efforts. It is now increasingly clear that even in case of genomes analysed using most advanced annotation pipelines, that underwent extensive manual curation of incorrect annotations resulting from sequencing errors, current annotations are far from complete even in well studied organisms like *Escherichia coli, Salmonella* Typhimurium *and Bacillus subtilis* (Baek et al. 2017; Giess et al. 2017; Meydan et al. 2019; Ndah et al. 2017). A plethora of data is available to illustrate these shortcomings across the phylogeny. In *Mycobacterium smegmatis*, a proteogenomic study identified 63 novel proteins and found evidence of upstream translation for 81 additional genes (Potgieter et al. 2016). Similarly, 30 novel genes and 50 erroneously annotated N-termini were identified in the plant pathogen *Xanthomonas euvesicatoria* (Abendroth et al. 2017). Transcription start site detection and proteomic validation also revealed 107 novel and 178 incorrectly assigned protein start sites in *Bradyrhizobium japonicum* (Cuklina et al. 2016*)*. Finally, in the same bacterial species, Kumar and co-workers identified 59 novel and 49 incorrectly annotated genes using the GenoSuite tool facilitating the proteogenomic analysis (Kumar et al. 2013).

Alternative translation initiation increases bacterial proteome complexity

The precise delineation of translation start sites is a persistent problem in prokaryotic genome annotation (Haft et al. 2018; Overmars et al. 2015). Importantly, the very relationship between a prokaryotic gene and protein needs to be revisited, since a single gene can give rise to multiple protein products, called proteoforms (Figure 1) (Gawron et al. 2014; Smith et al. 2013). It has previously been shown that up to 60% of annotated bacterial genomes contain incorrectly or non-assigned translation initiation sites (Nielsen and Krogh 2005), either due to misannotation or the expression of proteoforms. This complexity remains poorly captured in current prokaryotic genome annotations. It has been estimated that up to 10% of currently annotated bacterial proteins display

incorrect N-termini (Sato and Tajima 2012). Assignment of translation initiation sites (TIS) poses particular challenges because most open reading frames (ORFs) contain numerous in-frame start codons (Meydan et al. 2018). Next to the canonical AUG start codons, other "near-cognate" codons can serve as translation initiation sites in bacteria. More specifically, from available plasmid and genome sequences of ~70 bacteria, 18.2% of annotated start codons were non-AUG codons (Hecht et al. 2017). Such events often include their discovery in case of essential genes involved in DNA replication and conserved ribosomal proteins as demonstrated by *Deinococcus radiodurans and Deinococcus geothermalis* proteomics evidence (Baudet et al. 2010). A recent study further revealed translation initiation potential from 17 non-AUG codons in *E. coli*, predominantly GUG, UUG and to lesser extend CUG, AUU, AUA and AUC, which all may serve as near-cognate start codons decoded to methionine (Hecht et al. 2017). To preserve most coding information, typically only the longest ORF is annotated (Tatusova et al. 2016), which does not necessarily reflect the actual product translated *in vivo*. Moreover, different in-frame translation initiation sites upstream or downstream of the annotated coding sequence (CDS) may be differentially selected, leading to the expression of alternative (i.e. non-annotated) N-terminal (Nt-) proteoforms (Li et al. 2012; Oh et al. 2011). The few Nt-proteoforms reported in literature were usually discovered by coincidence during production and characterization of purified bacterial proteins (Miller and Wahba 1973; Nakagawa and Matsuhashi 1982). While sharing the same gene-origin, these protein variants display distinct but non-proteolytic protein N-termini which may impact protein functionality. Interestingly, the expression of (alternative) N-terminal proteoforms have already been implicated in regulation of virulence (by *ssaQ* (Yu et al. 2011)), photosynthesis (*ccmM* (Long et al. 2007), *petH* (Thomas et al. 2006)), response to heat shock (*clpB* (Chow and Baneyx 2005; Park et al. 1993)), translation (*infB* (Plumbridge et al. 1985)) and antibiotic production (*pikIV* (Xue and Sherman 2000)). Although precise functional studies investigating the biological function of such proteoforms are generally lacking, *E.coli* ClpB has been shown to facilitate recovery from heat shock if both, full length (857 aa) and N-terminally truncated (709 aa), proteoforms are coexpressed (Nagy et al. 2010). Further, several proteoforms raised upon alternative translation initiation have been shown to reside in very specific locations reflecting their unique functional properties. The N-terminal truncated SsaQ proteoform, crucial for *Salmonella* virulence, is found at the sorting base platform of a type III secretion system (Yu et al. 2011) while IF2-1 and IF2/3 proteoforms acting as *Escherichia* coli translation initiation factors are found in direct association with ribosomes (Caserta et al. 2006). Nt-proteoforms may lose or gain targeting signals and several such proteoforms were reported to reside in different subcellular locations as a consequence of alternative translation initiation (Cota-Gomez et al. 1997; Hwang et al. 1997; Mo et al. 1998). Taken together, comprehensive identification, annotation and characterization of such Nt-proteoforms is of key importance.

Further, out-of-frame start sites may be used for translation initiation, thus one mRNA molecule may potentially encode several (partially) overlapping ORFs. More specifically, a third of annotated bacterial gene structures are overlapping (Huvet and Stumpf 2014), the grand majority of which are oriented in the same direction and overlap with only four nucleotides (Johnson and Chisholm 2004). In case a downstream start and upstream stop codon are closely spaced in a polycistronic mRNA, gene translation can be coupled to translation of the respective upstream gene in a process known as re-initiation. In case of overlapping stop/start codons, re-initiation was found to be most efficiently coupled (Osterman et al. 2013). Despite the ubiquitous nature of such short overlaps, gene annotation pipelines typically do not consider more extensive overlapping sequences, leading to an underrepresentation of significantly overlapping gene structures, referred to as dually decoded regions (Michel et al. 2012), while nonetheless their occurrence have previously been reported in bacteria (Hucker et al. 2018). Next to overlapping on the same genomic strand, protein-coding ORFs can be located antisense (Konstantopoulou et al. 1995). Complementary evidence obtained by means of mass spectrometry and ribosome profiling or Ribo-seq, a method designed to measure genome-wide protein translation by sequencing of mRNA fragments occupied by translating ribosomes, provided evidence for the existence of small proteins encoded by small ORFs (sORFs) translated from RNAs antisense to annotated protein-coding genes in *Helicobacter pylori* (Friedman et al. 2017). In line, a recent ribosome profiling study presented experimental evidence for translation of eight novel small ORFs detected antisense to protein-coding genes in *E. coli* (Weaver et al. 2019). Just as in case of Nt-proteoforms, future riboproteogenomics efforts will aid in further deciphering the complexity of overlapping gene structures (Sberro et al. 2019).

So far we discussed protein complexity that depends on translation initiation leading to N-terminal proteoforms and translation of proteins encoded by overlapping ORFs. Alternative translation events such as ribosomal frameshifting, translational bypassing, stop codon read-through and codon reassignment can also result in expression of alternative gene translation products. Collectively these processes are referred to as alternative decoding strategies and have been comprehensively discussed in a recent review (Baranov et al. 2015). With currently employed genome annotation strategies however, capturing such translation events remains a considerable challenge, as also evident from recent proteogenomics efforts (Willems *et al.*, submitted, https://doi.org/10.1101/2019.12.18.881375; (Zheng et al. 2017)). In a striking example, the gene encoding release factor 2 (RF2), requiring +1 programmed frameshifting to produce the protein, has consistently been misannotated in majority of bacterial genomes (Bekaert et al. 2006). More recently, similar mechanism has been discovered in case of *E.coli* metal efflux pump CopA translation (Meydan et al. 2017). Such findings indicate missing annotations of frameshifting events even in,

unarguably, well characterized bacterial species. However, to fully facilitate the ribosome profiling aided discovery of bacterial frameshifting events, improvements to current bacterial ribosome profiling protocols are deemed necessary (Mohammad et al. 2019). More specifically, triplet periodicity (phasing) is a feature observed when ribosome profiling reads are aligned to single nucleotides, representing the location of the ribosomal P-sites. This feature is used to facilitate detection of the frame of translation. As compared to eukaryotic Ribo-seq and largely due to the use of a less specific nuclease to obtain bacterial footprints, bacterial Ribo-seq data generally has lower resolution, poor triplet periodicity and may suffer from sequence biases observed at footprint ends (Mohammad et al. 2019). These features significantly complicate ORF delineation in prokaryotes and further hinder the identification of multiple expressed Nt-proteoforms using elongating ribosome profiles alone, unless a downstream proteoform is notably more efficiently translated (Schrader et al. 2014).

Exploiting ribosome profiling for conditional bacterial genome (re-)annotation

Genomes are periodically curated and re-annotated once new experimental evidence and improved tools become available. RNA sequencing (RNA-seq), ribosome profiling, mass spectrometry (MS) data and phylogenetic analyses add complementary levels of information on top of the genome sequence and inform on the search for transcribed and translated genomic regions.

Recently, Ribo-seq revolutionized the annotation of protein-coding genomic regions and quantification of ORF translation (Ingolia et al. 2009). Ribo-seq was applied for the first time in bacteria by Oh *et al.* in 2011 (Oh et al. 2011). In bacteria, elongating ribosomes can be immobilized by flash freezing or pre-treatment with translation inhibitors, such as chloramphenicol. Subsequently, recovered ribosome protected fragments of mRNA are sequenced providing a snapshot of ongoing translation. The density of retrieved ribosome footprints report on translation efficiencies when compared to corresponding RNA-levels (Li et al. 2014) and distribution of footprints may show an accumulation at translation start and stop regions, aiding delineation of ORF boundaries (Giess et al. 2017; Ndah et al. 2017; Oh et al. 2011).

Several tools have been developed to predict translated bacterial ORFs (Giess et al. 2017; Ndah et al. 2017) using a combination of Ribo-seq metrics and mRNA sequence features (e.g. Shine Dalgarno (SD) ribosome binding site (Shine and Dalgarno 1975)). Applying an integrative algorithm, termed REPARATION (Ndah et al. 2017), to *Salmonella enterica* serovar Typhimurium*, Escherichia coli* and *Bacillus subtilis* Ribo-seq data, we predicted 3421, 3202 and 3435 expressed ORFs, respectively. Among these ORFs, 501, 249 and 477 pointed to the expression of extended or truncated Nt-proteoforms, including the highly conserved 5’ extension of adhP in the *Salmonella* strain SL1344

(Figure 2). Additionally, in the three species analyzed, we identified 98, 98 and 225 novel ORFs. These novel putative coding regions were more frequently initiated at near-cognate start sites than annotated ORFs but had very comparable amino-acid composition and between 19 and 59% had at least one conserved orthologue in a related species. Overall, applying the algorithm to *Salmonella* strain SL1344 revealed nearly 600 proteins with an incorrect or missing annotation, albeit genome alignment using MAUVE (Darling et al. 2004) revealed that 8% of them were correctly annotated in the related *Salmonella* strain 14028S and in *E. coli* K-12 (Figure 2). Due to the Ribo-seq limitations mentioned above, in REPARATION only a single, most abundantly expressed and uniformly covered ORF is selected per identified ORF family, again introducing a bias against the discovery of multiple expressed proteoforms. This limitation was partially addressed by generating a corresponding N-terminal proteomics dataset, leading to the univocal identification of 11 *E. coli* genes with at least two selected TIS (Ndah et al. 2017).

Besides REPARATION, an alternative approach for ORF delineation was developed with a particular focus on TIS prediction (Giess et al. 2017). By investigating ribosome P-site assignment in relation to ribosome footprint length, we could show that consideration of read length distributions around translation initiation sites aided precise TIS delineation. Overall, extracted features combined with Ribo-seq metrics, and TIS context information (in particular anti-SD interaction) were used to train a random forest algorithm. Using elongating Ribo-seq data, 214 extensions, 205 truncations and 61 novel ORFs were discovered in *Salmonella.* Notably, this model successfully handled initiating ribosome data from *E. coli* (Nakahigashi et al. 2016), allowing the discovery of 11 extensions and 13 truncations*.* More recently, a new tool named DeepRibo provided increased performance in ORF detection based on deep learning neural networks (Clauwaert et al. 2019). With the increasing availability of Ribo-seq datasets from diverse microorganisms, we can expect that prokaryotic (translated) ORF prediction tools will continue to improve and will generally be adopted for ORF (re-)annotation.

In this context, recent advances in ribosome profiling of prokaryotes using different translation inhibitors definitely aid the annotation process. In contrast to chloramphenicol, which inhibits elongation through the peptidyl transferase center, tetracycline prevents tRNA binding directly at ribosome A-site (Wilson 2009). Tetracycline-inhibited ribosome profiling (TetRP) provided footprint profiles with distinctive accumulation of ribosomes around start site (Nakahigashi et al. 2014; Nakahigashi et al. 2016). Although both chloramphenicol and tetracycline block elongation and exhibit footprint coverage across ORF body, TetRP produced sharper TIS peaks, facilitating genome-wide detection of translation start sites in *E. coli* (Nakahigashi et al. 2016). Using TetRP, 154 genes

with alternative TIS were discovered, revealing many missed or misannotated start sites in the *E. coli* genome. Interestingly, 28 genes encoding two different N-terminal proteoforms were identified.

However, it is only with the recent application of retapamulin, a small molecule drug interfering with peptidyl transferase center of the ribosome, that a real breakthrough for translation initiation profiling in bacteria was made (Meydan et al. 2019). Retapamulin, as well as another tested inhibitor Onc112 (a proline-rich antimicrobial peptide that binds in the exit tunnel and prohibits tRNA binding at ribosome A-site), offered superior resolution for start codon identification than tetracycline (Weaver et al. 2019), and is arguably superior to lactimidomycin and harringtonine assisted Ribo-seq in eukaryotes (Ingolia et al. 2011; Lee et al. 2012), further improving P-site assignment. Retapamulin-assisted Ribo-seq data (Ribo-RET) confirmed 991 annotated TIS from 1153 expressed genes in the *E. coli* BW25113 strain (Meydan et al. 2019). Additionally, 124 internal TIS in relation to annotated ORFs were discovered in two *E. coli* stains (BW25113 and BL21, Figure 3A). More specifically, 42 in-frame and 74 out-of-frame TIS pointed to the expression of N-terminally truncated proteoforms and out-of-frame ORF translation products, respectively. Besides, 8 retapamulin peaks were detected at non-cognate codons (i.e. non-AUG, GUG, UUG, CUG or AUU codons or so-called “non-start” codons).

Nowadays, with the plethora of available Ribo-seq datasets, translation of a gene of interest can easily be inspected. Currently, publically available data from a variety of bacterial species, including *Escherichia coli*, *Bacillus subtilis*, *Caulobacer crescentus*, *Streptomyces coelicolor*, *Staphylococcus aureus*, *Mycobacterium abscessus*, *Salmonella* Typhimurium and *Pseudomonas aeruginosa* can be accessed via the GWIPS-viz genome browser (http://gwips.ucc.ie/)(Michel et al. 2014)). Interestingly, when performing a survey of Ribo-seq data compiled by GWIPS-viz, and focusing on alternative TISs identified by means of Ribo-RET (Meydan et al. 2019), patterns of differential and conditional proteoform expression in *E.coli* could be revealed (Figure 4).

Bacterial genome (re-)annotations informed by proteomics

MS-driven proteomics provides the ultimate evidence for protein expression. Whole-proteome studies however, face challenges in comprehensive proteoform discovery, especially when it comes to the unambiguous identification of proteoform-specific peptides. It has long been proposed that each annotation effort should be accompanied by a cost-efficient interrogation of the respective proteome (Gupta et al. 2007), and in the contect of proteogenomics, numerous proteomic approaches have been used to facilitate annotation efforts. Next to shotgun proteomics, MALDI analysis also enabled the detection of 7 new gene products in *Shigella flexneri* (Zhao et al. 2011), clearly indicating the omnipresence of alternative and unannotated proteoform expression. However, Nt-proteoforms typically share the majority of identifiable peptides with their annotated

counterparts and often can solely be discriminated by their N-terminal peptides (Berry et al. 2016; Koch et al. 2014).

Therefore, proteoforms generated by alternative translation initiation can efficiently be captured and identified by enriching their N-terminal peptides prior to LC-MS/MS analysis. High-throughput N-terminal proteomics (i.e. N-terminomics) studies were originally made possible with the introduction of N-terminal COFRADIC (COmbined FRActional DIagonal Chromatography) (Gevaert et al. 2003; Staes et al. 2011). Different N-terminal proteomics technologies (Marino et al. 2015) and their applicability for the identification of (alternative) protein N-termini in bacteria have already been reviewed (Berry et al. 2016). For bacterial protein synthesis, N-formylmethionine (fMet) is specifically used for translation initiation. Subsequently, during translation, prokaryotic protein N-termini can be co-translationally modified (Figure 1). Both properties are helpful in distinguishing true translation-indicative N-termini from N-termini raised upon proteolysis. First, peptide deformylase (PDF) nearly quantitatively removes the formyl group from the initiator methionine (iMet), thereby exposing a free N-terminal amine (Meinnel et al. 1993). Furthermore, deformylation is essentially required for subsequent iMet excision to occur (Solbiati et al. 1999). Dependent on the identity of the second amino acid residue, iMet is removed by methionine aminopeptidases (MetAPs) in approximately half of mature bacterial proteins (Frottin et al. 2006). Subsequently, N-termini may be subjected to (partial) Nt-acetylation by N-terminal acetyl transferases (NATs), a process which occurs post-translationally in bacteria (VanDrisse and Escalante-Semerena 2019). Of note however, while high degrees of bacterial protein Nt-acetylation are rare, low degrees (lower than 10%) have recently been reported for a handful of bacterial proteins (Bienvenut et al. 2015). Deformylation may be incomplete for some N-termini (e.g. internal membrane proteins with N-out topology (Ranjan et al. 2017)) or chemically inhibited using PDF inhibitors (e.g. actinonin (Bienvenut et al. 2015; Impens et al. 2017). Consequently, identification of Nt-formylated or Nt-acetylated peptides provides the ultimate evidence for translation initiation events (Giess et al. 2017; Impens et al. 2017). Additionally, for the comprehensive identification of (deformylated) N-termini, various N-terminomics technologies include mass tagging of primary amines before protein digestion to univocally enable the detection of *in vivo* proteoform starts. TIS-indicative N-terminal peptides can subsequently be verified for compliance with known rules of iMet processing by MetAPs and presence of other corresponding genome features, such as a Shine-Dalgarno sequence. Altogether, such strategies provide powerful tools to study (alternative) translation, thereby aiding the (re-) annotation of bacterial genomes (Ansong et al. 2011; Bland et al. 2014; Giess et al. 2017; Impens et al. 2017; Ndah et al. 2017; Van Damme et al. 2014). Further, MS-based peptide identification critically depends on comprehensive protein sequence databases. In other words, only peptides included in the search

database can be identified. In some cases, however, the mere presence in the protein database may still be insufficient. Although proteoform sequences starting at internal methionine are fully embedded inside annotated sequences, their N-terminal peptides are semi-tryptic (i.e. the true N-terminus of the protein is not a result of *in silico* trypsin cleavage) and can thus only be identified by adjusting the enzymatic search parameters accordingly. N-terminal proteoforms initiated at non-AUG codons and decoded to methionine (Van Damme et al. 2014) pose an additional difficulty, because of the apparent mismatch between nucleotide and protein sequence.

Using N-terminomics and six-frame stop-to-stop genome translation of the marine bacterium *Roseobacter denitrificans*, Bland *et al.* validated 534 N-termini, including 41 N-termini indicative of misannotated start codons, besides the identification of N-termini which mapped to 5 novel intergenic ORF translation products and to 8 genes with more than one Nt-proteoform (Bland et al. 2014). Impens *et al.* applied N-terminal COFRADIC to study the proteoform repertoire of *L. monocytogenes* (Impens et al. 2017). A six-frame translation database search complemented with another smaller, rationalized protein sequence search space containing (N-terminally extended) variants of canonical proteins, enabled the identification of so-called 5' untranslated region (5'UTR) translation products, next to translation products of putative ORFs in short RNAs and antisense transcripts. Overall, this approach revealed 19 truncated Nt-proteoforms and 6 novel miniproteins. Using a similar N-terminomics strategy, Bienvenut *et al.* (Bienvenut et al. 2015) identified 1228 distinct N-termini in *E. coli* and characterised a subset of 510 N-termini over time upon PDF inhibition. This data revealed patterns of N-terminal modifications originating from PDF, MetAP and NAT activities which led to the discovery of 11 truncated Nt-proteoforms. Moreover, proteogenomic analysis of *Meyhylbacterium extorquens* N-terminome revealed 39 new proteins and uncovered 78 incorrectly assigned translation start sites (Bibi-Triki et al. 2018). More recently, we demonstrated that, by using an optimized proteogenomic workflow, an enhanced proteome-depth and greater confidence for unannotated peptide hits was achieved (Willems et al., submitted, https://doi.org/10.1101/2019.12.18.881375). Overall, these aforementioned studies clearly highlight that experimental data assisted genome annotation is of indispensable value for the correction of annotation errors.

An alternative strategy in proteomics-assisted genome (re-)annotation is the integration of transcriptome and/or translatome evidence in the process of customized protein database construction. We pioneered the systematic implementation of ribo-seq data into proteogenomic studies of eukaryotes (Crappe et al. 2015; Verbruggen et al. 2019) and expanded this strategy to prokaryotes (Giess et al. 2017; Ndah et al. 2017). Alternative translation frames delineated from ribosome profiling data were considered and subsequently transformed into customized, sample-

oriented sequence databases to serve as reference for the proteomes under study. N-terminomics and shotgun proteomics of *E. coli* and *Salmonella* validated the expression of Ribo-seq predicted ORFs (Figure 2A). Relying on the detection of N-terminally modified peptides, an efficient discovery of proteoforms was enabled. Importantly, this approach helped to overcome limitations in identifying expressed Nt-proteoform pairs, including the identification of an Nt-proteoform pair encoded by *E. coli grpE*, showing two translation initiation sites spaced by only 19 codons (Ndah et al. 2017). A similar approach was presented by Nakahigashi *et al.* (Nakahigashi et al. 2016), who used N-terminal proteomics to validate TetRP data, confirming 5 genes giving rise to multiple Nt-proteoforms via alternative translation initiation.

Functional outcome of alternative proteoforms

We and others identified numerous (conserved) N-terminal proteoforms (N-terminal truncations or extensions of proteins) in a variety of bacterial species (Baek et al. 2017; Clauwaert et al. 2019; Giess et al. 2017; Impens et al. 2017; Meydan et al. 2019; Nakahigashi et al. 2016; Ndah et al. 2017), overall revealing a more complete picture of bacterial proteome complexity. Consequently, newly identified bacterial proteoforms await further characterization. We analysed the translation products of in-frame internal initiation events captured by retapamulin-assisted Ribo-seq (Meydan et al. 2019). Interestingly, we found that the resulting truncated proteoforms were generally less hydrophobic (adj. p-value = 8.8e-3, Figure 3B) but more disordered (adj. p-value = 7.3e-4, Figure 3C) as compared to their full-length annotated counterparts. More specifically, when proteoform pairs (full length vs. truncated) were considered for the 42 abovementioned genes, 7 truncated Nt-proteoforms lost a transmembrane (TM) domain according to Phobius (Kall et al. 2004) and/or TOPCONS (Tsirigos et al. 2015) predictions (Figure 5A). Of note, for 18 of the 42 internal TIS identified, retapamulin signal was additionally observed at the corresponding database annotated TIS (dbTIS, Figure 5B), indicative of the expression of both proteoforms. Additionally, in case of *fabB, slyB, mlaD, dsbG* and *yadE* gene translation, SignalP predictions (Almagro Armenteros et al. 2019) revealed the expression of an Nt-truncated proteoform that lost its N-terminal signal peptides (SP, Figure 5A,C). When inspecting in-house generated Retapamulin-assisted Ribo-seq data, signal intensity at start sites was found to correlate well with translation efficiency and protein abundance (Pearson correlation coefficient $r = 0.68$ and $r = 0.56$, respectively, unpublished data). Assuming quantitative properties of retapamulin data, in case of 4 genes with both the full-length and SP-lacking Nt-proteoform identified, the truncated proteoform was repeatedly less expressed than its full-length counterpart (Figure 5B). However, truncated proteoforms with higher expression levels as compared to their annotated counterparts was also found in case of *infB, speA, arcB, yebG, bamA* (Figure 5B). Interestingly, an internal initiation (iTIS) in *mgtS* gene was predicted to reveal a cryptic signal peptide

in the corresponding 31 amino acid long Nt-truncated proteoform (Figure 5B, D). Overall these meta-analyses suggest that many of the Nt-proteoforms identified in *E.coli* - and by extension in bacteria in general - might display a differential (sub-)cellular localization.

Our predictions point to significant differences in proteoform subcellular targeting. Clearly, spatial aspects of protein expression cannot be overlooked as protein localization may provide further hints towards (differences in) proteoform functionality. Differential localisation of proteoforms was reported for several bacterial translation products which appeared to lose or gain N-terminal targeting sequences as a result of alternative translation initiation. As inferred from the examples described below, this mechanism has proven instrumental in directing protein activity to several subcellular niches without increasing the size of the bacterial genome. In *P. aeruginosa*, *plcR* regulates the secretion of a haemolytic enzyme phospholipase C (PlcH) (Cota-Gomez et al. 1997) and encodes two Nt-proteoforms resulting from alternative translation initiation. The longer proteoform is localised in the periplasm and a shorter, lacking the signal peptide, was retained in the cytosol. These differentially localised Nt-proteoforms may ensure proper PlcH trafficking, one acting as cytoplasmic chaperone, the other as PlcH carrier between the inner and outer membranes (Cota-Gomez et al. 1997). In *E. coli*, the plasmid-encoded CvaA protein is required for colicin V secretion. Next to the canonical, membrane-associated CvaA proteoform, a soluble, truncated proteoform lacking the N-terminal transmembrane stretch is expressed (Hwang et al. 1997). The short proteoform protects the longer one from degradation thereby enhancing colicin V secretion. Another interesting example of differentially localised proteoforms was reported in *Coxiella brunetii*. This bacterial parasite with reduced genome has only one copy of the *mip* virulence gene (i.e. *cbmip*) instead of the multiple *mip* copies typically expressed in species like *Legionella* and *Chlamydia*. In this bacterium however, targeting of Mip peptidyl-prolyl isomerase activity to different subcellular compartments is achieved via CbMip translation from 3 alternative TIS (Mo et al. 1998). Similarly, an antibacterial toxin lactococcin A of *Lactococcus lactis* is secreted using the LcnC and D proteins. It has been shown that lcnD gene produces two proteoforms, with longer displaying membrane localization and shorter residing in the cytoplasm (Varcamonti et al. 2001).

Although many subcellular protein localization maps were drafted in eukaryotes (Christoforou et al. 2016; Geladaki et al. 2019; Itzhak et al. 2016; Itzhak et al. 2017), in prokaryotes only limited data with lower resolution is available in few model organisms (Brown et al. 2010; Brown et al. 2012; Fontaine et al. 2011; Ohniwa et al. 2011; Omasits et al. 2013; Orfanoudaki and Economou 2014; Pieper et al. 2009; Stekhoven et al. 2014; Thein et al. 2010). Spatial mapping of proteins often relies on the combination of subcellular fractionation with proteomics readout enabling the determination of protein localisations. As such, several recent studies applied ultracentrifugation and differential

solubility in sarkosyl (1%) followed by shotgun proteomics to discriminate cytosolic, periplasmatic, inner and outer membrane (-associated) proteins, thereby creating a subcellular map of 1000 *Salmonella* proteins (Brown et al. 2012) and 1025 *Bartonella henselae* proteins (Omasits et al. 2013; Stekhoven et al. 2014) expressed under standard and infection-relevant conditions. An important advantage of subcellular fractionation was the increased bacterial proteome coverage largely attributed to the identification of low abundant or hydrophobic proteins, protein categories typically underrepresented in total proteome studies (Brown et al. 2010; Brown et al. 2012; Impens et al. 2017; Yuan et al. 2018). In light of these findings, comprehensive mapping of protein localizations in various bacteria and across growth conditions, combined with customised proteoform databases for searching mass spectrometry data and N-terminomics, could shed more light on spatial distribution and functional properties of Nt-proteoforms.

Besides localisation, studying protein synthesis and degradation rates can further elucidate relevant properties of novel proteoforms. Protein synthesis and stability are partially determined by sequence features, namely codon composition and (N-terminal) amino acid identity, which is particularly important in the context of Nt-proteoforms. Protein-coding ORFs adapted their synonymous codon frequency to the availability of tRNA pools (Ikemura 1981) and ribosomal translation dynamics (Boel et al. 2016), thereby increasing protein yield *in vivo*. In consequence, codon composition of ORFs divergent from annotated genes, may exert negative influence on their expression efficiency (Puigbo et al. 2008). While codons impact protein synthesis, N-terminal amino acids and their modifications impact protein degradation (Apel et al. 2010; Piatkov et al. 2015; Tobias et al. 1991). According to N-end rules, certain destabilizing residues exposed as protein N-termini (R, K, L, F, Y, W absent from natural N-termini, rather derived from proteolysis), as well as retained formyl-methionines, act as signals for protein degradation by the proteasome-like ClpAP protease (Piatkov et al. 2015; Tobias et al. 1991). Indeed, the steady-state levels of formylated N-termini detected in *Salmonella* and *E. coli* are very low (Giess et al. 2017; Ndah et al. 2017). The identity of the penultimate N-terminal residue combined with the activity of MetAPs also contribute to the regulation of bacterial protein stability (Apel et al. 2010), a finding in agreement with our previous observations of altered Nt-proteoform stability in human (Gawron et al. 2016). We previously demonstrated that human Nt-proteoforms of one gene may have dramatically different stability, even when missing just one or a few amino acids at the N-terminus (Gawron et al. 2016). Future studies are needed to reveal if such regulation exists in bacteria.

The ultimate evidence for functional diversification of proteoforms requires in-depth molecular and biochemical studies. So far, several alternative proteoforms were shown to mediate unique protein-

protein interactions and act differently during complex formation. The long proteoform of CheA is a kinase of the *E. coli* signalling cascade regulating chemotaxis. The truncated CheA proteoform interacts with both phosphorylating and dephosphorylating protein complexes of this pathway (Wang and Matsumura 1997), suggesting a regulatory role of this proteoform. In cyanobacteria, CcmM proteoforms are involved in Rubisco complex assembly during carbon dioxide assimilation in specialized microcompartments called carboxysomes. The Nt-truncated proteoform of CcmM interlinks Rubisco enzymes, while the full-sized CcmM anchors this complex to the carboxysome lumen (Long et al. 2007). In *Salmonella*, coexpression of SsaQ proteoforms is necessary for the formation of the type III secretion system (Yu et al. 2011) allowing these bacteria to inject effectors into infected cells. Future mapping of proteoform-specific protein interactors may be steered by applying high-throughput technologies. Ribosome profiling data has shown that translation efficiency is proportional to the stoichiometry of components of several large multi-member protein complexes (Li et al. 2014). In order to obtain comprehensive interactomics data to complement Ribo-seq-derived translation efficiency information, one may employ endogenous tagging followed by immunocapture or alternatively, proximity ligation (Branon et al. 2018) coupled to MS. Point mutations of TIS codons could be used to steer (the absence of) proteoform expression and can thus be applied to distinguish and compare the interactomes of Nt-proteoform pairs and to elucidate certain proteoform-specific phenotypic parameters including susceptibility to antibiotics, growth rate and infectivity potential in case of pathogenic bacteria. Overall, large-scale proteoform discovery and characterisation will complement ongoing efforts to curate prokaryotic genome annotation next to increasing our understanding of bacterial proteoform biology.

References

Abendroth, U., et al. (2017), 'Identification of new protein-coding genes with a potential role in the virulence of the plant pathogen Xanthomonas euvesicatoria', *BMC Genomics,* 18 (1), 625.

Almagro Armenteros, J. J., et al. (2019), 'SignalP 5.0 improves signal peptide predictions using deep neural networks', *Nat Biotechnol,* 37 (4), 420-23.

Ansong, C., et al. (2011), 'Experimental annotation of post-translational features and translated coding regions in the pathogen Salmonella Typhimurium', *BMC Genomics,* 12, 433.

Apel, W., et al. (2010), 'Identification of protein stability determinants in chloroplasts', *Plant J,* 63 (4), 636-50.

Baek, J., et al. (2017), 'Identification of Unannotated Small Genes in Salmonella', *G3 (Bethesda),* 7 (3), 983-89.

Baranov, P. V., et al. (2015), 'Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning', *Nat Rev Genet,* 16 (9), 517-29.

Baudet, M., et al. (2010), 'Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons', *Mol Cell Proteomics,* 9 (2), 415-26.

Bekaert, M., et al. (2006), 'ARFA: a program for annotating bacterial release factor genes, including prediction of programmed ribosomal frameshifting', *Bioinformatics,* 22 (20), 2463-5.
Berry, I. J., et al. (2016), 'The application of terminomics for the identification of protein start sites and proteoforms in bacteria', *Proteomics,* 16 (2), 257-72.
Bibi-Triki, S., et al. (2018), 'N-terminome and proteogenomic analysis of the Methylobacterium extorquens DM4 reference strain for dichloromethane utilization', *J Proteomics,* 179, 131-39.
Bienvenut, W. V., et al. (2015), 'Proteome-wide analysis of the amino terminal status of Escherichia coli proteins at the steady-state and upon deformylation inhibition', *Proteomics,* 15 (14), 2503-18.
Bland, C., et al. (2014), 'N-Terminal-oriented proteogenomics of the marine bacterium roseobacter denitrificans Och114 using N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography', *Mol Cell Proteomics,* 13 (5), 1369-81.
Blattner, F. R., et al. (1997), 'The complete genome sequence of Escherichia coli K-12', *Science,* 277 (5331), 1453-62.
Boel, G., et al. (2016), 'Codon influence on protein expression in E. coli correlates with mRNA levels', *Nature,* 529 (7586), 358-63.
Branon, T. C., et al. (2018), 'Efficient proximity labeling in living cells and organisms with TurboID', *Nat Biotechnol,* 36 (9), 880-87.
Brown, R. N., et al. (2010), 'Mapping the subcellular proteome of Shewanella oneidensis MR-1 using sarkosyl-based fractionation and LC-MS/MS protein identification', *J Proteome Res,* 9 (9), 4454-63.
Brown, R. N., et al. (2012), 'A Comprehensive Subcellular Proteomic Survey of Salmonella Grown under Phagosome-Mimicking versus Standard Laboratory Conditions', *Int J Proteomics,* 2012, 123076.
Caserta, E., et al. (2006), 'Translation initiation factor IF2 interacts with the 30 S ribosomal subunit via two separate binding sites', *J Mol Biol,* 362 (4), 787-99.
Chow, I. T. and Baneyx, F. (2005), 'Coordinated synthesis of the two ClpB isoforms improves the ability of Escherichia coli to survive thermal stress', *FEBS Lett,* 579 (20), 4235-41.
Christoforou, A., et al. (2016), 'A draft map of the mouse pluripotent stem cell spatial proteome', *Nat Commun,* 7, 8992.
Clauwaert, J., et al. (2019), 'DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns', *Nucleic Acids Res,* 47 (6), e36.
Cota-Gomez, A., et al. (1997), 'PlcR1 and PlcR2 are putative calcium-binding proteins required for secretion of the hemolytic phospholipase C of Pseudomonas aeruginosa', *Infect Immun,* 65 (7), 2904-13.
Crappe, J., et al. (2015), 'PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration', *Nucleic Acids Res,* 43 (5), e29.
Cuklina, J., et al. (2016), 'Genome-wide transcription start site mapping of Bradyrhizobium japonicum grown free-living or in symbiosis - a rich resource to identify new transcripts, proteins and to study gene regulation', *BMC Genomics,* 17, 302.
Darling, A. C., et al. (2004), 'Mauve: multiple alignment of conserved genomic sequence with rearrangements', *Genome Res,* 14 (7), 1394-403.
Finn, R. D., et al. (2016), 'The Pfam protein families database: towards a more sustainable future', *Nucleic Acids Res,* 44 (D1), D279-85.
Fontaine, F., et al. (2011), 'Membrane localization of small proteins in Escherichia coli', *J Biol Chem,* 286 (37), 32464-74.
Friedman, R. C., et al. (2017), 'Common and phylogenetically widespread coding for peptides by bacterial small RNAs', *BMC Genomics,* 18 (1), 553.
Frottin, F., et al. (2006), 'The proteomics of N-terminal methionine cleavage', *Mol Cell Proteomics,* 5 (12), 2336-49.

Gawron, D., et al. (2014), 'The proteome under translational control', *Proteomics,* 14, 2647-62.
Gawron, D., et al. (2016), 'Positional proteomics reveals differences in N-terminal proteoform stability', *Mol Syst Biol,* 12 (2), 858.
Geladaki, A., et al. (2019), 'Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics', *Nat Commun,* 10 (1), 331.
Gevaert, K., et al. (2003), 'Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides', *Nat Biotechnol,* 21 (5), 566-9.
Giess, A., et al. (2017), 'Ribosome signatures aid bacterial translation initiation site identification', *BMC Biol,* 15 (1), 76.
Gupta, N., et al. (2007), 'Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation', *Genome Res,* 17 (9), 1362-77.
Haft, D. H., et al. (2013), 'TIGRFAMs and Genome Properties in 2013', *Nucleic Acids Res,* 41 (Database issue), D387-95.
Haft, D. H., et al. (2018), 'RefSeq: an update on prokaryotic genome annotation and curation', *Nucleic Acids Res,* 46 (D1), D851-D60.
Hecht, A., et al. (2017), 'Measurements of translation initiation from all 64 codons in E. coli', *Nucleic Acids Res,* 45 (7), 3615-26.
Hooda, Y., et al. (2017), 'Identification of a Large Family of Slam-Dependent Surface Lipoproteins in Gram-Negative Bacteria', *Front Cell Infect Microbiol,* 7, 207.
Hucker, S. M., et al. (2018), 'A novel short L-arginine responsive protein-coding gene (laoB) antiparallel overlapping to a CadC-like transcriptional regulator in Escherichia coli O157:H7 Sakai originated by overprinting', *BMC Evol Biol,* 18 (1), 21.
Huvet, M. and Stumpf, M. P. (2014), 'Overlapping genes: a window on gene evolvability', *BMC Genomics,* 15, 721.
Hwang, J., et al. (1997), 'Characterization of in-frame proteins encoded by cvaA, an essential gene in the colicin V secretion system: CvaA* stabilizes CvaA to enhance secretion', *J Bacteriol,* 179 (3), 689-96.
Hyatt, D., et al. (2010), 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics,* 11, 119.
Ikemura, T. (1981), 'Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes', *J Mol Biol,* 146 (1), 1-21.
Impens, F., et al. (2017), 'N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes', *Nat Microbiol,* 2, 17005.
Ingolia, N. T., et al. (2011), 'Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes', *Cell,* 147 (4), 789-802.
Ingolia, N. T., et al. (2009), 'Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling', *Science,* 324 (5924), 218-23.
Itzhak, D. N., et al. (2016), 'Global, quantitative and dynamic mapping of protein subcellular localization', *Elife,* 5.
Itzhak, D. N., et al. (2017), 'A Mass Spectrometry-Based Approach for Mapping Protein Subcellular Localization Reveals the Spatial Proteome of Mouse Primary Neurons', *Cell Rep,* 20 (11), 2706-18.
Jarvik, T., et al. (2010), 'Short-term signatures of evolutionary change in the Salmonella enterica serovar typhimurium 14028 genome', *J Bacteriol,* 192 (2), 560-7.
Johnson, Z. I. and Chisholm, S. W. (2004), 'Properties of overlapping genes are conserved across microbial genomes', *Genome Res,* 14 (11), 2268-72.
Kall, L., et al. (2004), 'A combined transmembrane topology and signal peptide prediction method', *J Mol Biol,* 338 (5), 1027-36.
Kelley, D. R., et al. (2012), 'Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering', *Nucleic Acids Res,* 40 (1), e9.

Koch, A., et al. (2014), 'A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites', *Proteomics,* 14 (23-24), 2688-98.
Konstantopoulou, I., et al. (1995), 'A Drosophila hsp70 gene contains long, antiparallel, coupled open reading frames (LAC ORFs) conserved in homologous loci', *J Mol Evol,* 41 (4), 414-20.
Kroger, C., et al. (2012), 'The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium', *Proc Natl Acad Sci U S A,* 109 (20), E1277-86.
Kumar, D., et al. (2013), 'Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline', *Mol Cell Proteomics,* 12 (11), 3388-97.
Lee, S., et al. (2012), 'Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution', *Proc Natl Acad Sci U S A,* 109 (37), E2424-32.
Li, G. W., et al. (2012), 'The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria', *Nature,* 484 (7395), 538-41.
Li, G. W., et al. (2014), 'Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources', *Cell,* 157 (3), 624-35.
Lloyd, C. R., et al. (2017), 'The Small Protein SgrT Controls Transport Activity of the Glucose-Specific Phosphotransferase System', *J Bacteriol,* 199 (11).
Lomsadze, A., et al. (2018), 'Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes', *Genome Res,* 28 (7), 1079-89.
Long, B. M., et al. (2007), 'Analysis of carboxysomes from Synechococcus PCC7942 reveals multiple Rubisco complexes with carboxysomal proteins CcmM and CcaA', *J Biol Chem,* 282 (40), 29323-35.
Marcellin, E., et al. (2013), 'Re-annotation of the Saccharopolyspora erythraea genome using a systems biology approach', *BMC Genomics,* 14, 699.
Marino, G., et al. (2015), 'Protein Termini and Their Modifications Revealed by Positional Proteomics', *ACS Chem Biol,* 10 (8), 1754-64.
McClelland, M., et al. (2001), 'Complete genome sequence of Salmonella enterica serovar Typhimurium LT2', *Nature,* 413 (6858), 852-6.
Meinnel, T., et al. (1993), 'Methionine as translation start signal: a review of the enzymes of the pathway in Escherichia coli', *Biochimie,* 75 (12), 1061-75.
Meydan, S., et al. (2018), 'Genes within Genes in Bacterial Genomes', *Microbiol Spectr,* 6 (4).
Meydan, S., et al. (2019), 'Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome', *Mol Cell,* 74 (3), 481-93 e6.
Meydan, S., et al. (2017), 'Programmed Ribosomal Frameshifting Generates a Copper Transporter and a Copper Chaperone from the Same Gene', *Mol Cell,* 65 (2), 207-19.
Michel, A. M., et al. (2012), 'Observation of dually decoded regions of the human genome using ribosome profiling data', *Genome Res,* 22 (11), 2219-29.
Michel, A. M., et al. (2014), 'GWIPS-viz: development of a ribo-seq genome browser', *Nucleic Acids Res,* 42 (Database issue), D859-64.
Miller, M. J. and Wahba, A. J. (1973), 'Chain initiation factor 2. Purification and properties of two species from Escherichia coli MRE 600', *J Biol Chem,* 248 (3), 1084-90.
Miravet-Verde, S., et al. (2019), 'Unraveling the hidden universe of small proteins in bacterial genomes', *Mol Syst Biol,* 15 (2), e8290.
Mo, Y. Y., et al. (1998), 'Synthesis in Escherichia coli of two smaller enzymically active analogues of Coxiella burnetii macrophage infectivity potentiator (CbMip) protein utilizing a single open reading frame from the cbmip gene', *Biochem J,* 335 (Pt 1), 67-77.
Mohammad, F., et al. (2019), 'A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution', *Elife,* 8.
Nagy, M., et al. (2010), 'Synergistic cooperation between two ClpB isoforms in aggregate reactivation', *J Mol Biol,* 396 (3), 697-707.

Nakagawa, J. and Matsuhashi, M. (1982), 'Molecular divergence of a major peptidoglycan synthetase with transglycosylase-transpeptidase activities in Escherichia coli --- penicillin-binding protein 1Bs', *Biochem Biophys Res Commun,* 105 (4), 1546-53.

Nakahigashi, K., et al. (2014), 'Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo', *BMC Genomics,* 15, 1115.

Nakahigashi, K., et al. (2016), 'Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling', *DNA Res,* 23 (3), 193-201.

Ndah, E., et al. (2017), 'REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes', *Nucleic Acids Res,* 45 (20), e168.

Nielsen, P. and Krogh, A. (2005), 'Large-scale prokaryotic gene prediction and comparison to genome annotation', *Bioinformatics,* 21 (24), 4322-9.

Oh, E., et al. (2011), 'Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo', *Cell,* 147 (6), 1295-308.

Ohniwa, R. L., et al. (2011), 'Proteomic analyses of nucleoid-associated proteins in Escherichia coli, Pseudomonas aeruginosa, Bacillus subtilis, and Staphylococcus aureus', *PLoS One,* 6 (4), e19172.

Omasits, U., et al. (2013), 'Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome', *Genome Res,* 23 (11), 1916-27.

Orfanoudaki, G. and Economou, A. (2014), 'Proteome-wide subcellular topologies of E. coli polypeptides database (STEPdb)', *Mol Cell Proteomics,* 13 (12), 3674-87.

Osterman, I. A., et al. (2013), 'Comparison of mRNA features affecting translation initiation and reinitiation', *Nucleic Acids Res,* 41 (1), 474-86.

Overmars, L., et al. (2015), 'A Novel Quality Measure and Correction Procedure for the Annotation of Microbial Translation Initiation Sites', *PLoS One,* 10 (7), e0133691.

Park, S. K., et al. (1993), 'Site-directed mutagenesis of the dual translational initiation sites of the clpB gene of Escherichia coli and characterization of its gene products', *J Biol Chem,* 268 (27), 20170-4.

Parsons, M., et al. (2015), 'Advancing Trypanosoma brucei genome annotation through ribosome profiling and spliced leader mapping', *Mol Biochem Parasitol,* 202 (2), 1-10.

Piatkov, K. I., et al. (2015), 'Formyl-methionine as a degradation signal at the N-termini of bacterial proteins', *Microb Cell,* 2 (10), 376-93.

Pieper, R., et al. (2009), 'Integral and peripheral association of proteins and protein complexes with Yersinia pestis inner and outer membranes', *Proteome Sci,* 7, 5.

Plumbridge, J. A., et al. (1985), 'Two translational initiation sites in the infB gene are used to express initiation factor IF2 alpha and IF2 beta in Escherichia coli', *EMBO J,* 4 (1), 223-9.

Poptsova, M. S. and Gogarten, J. P. (2010), 'Using comparative genome analysis to identify problems in annotated microbial genomes', *Microbiology,* 156 (Pt 7), 1909-17.

Potgieter, M. G., et al. (2016), 'Proteogenomic Analysis of Mycobacterium smegmatis Using High Resolution Mass Spectrometry', *Front Microbiol,* 7, 427.

Promponas, V. J., et al. (2015), 'Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure', *Stand Genomic Sci,* 10, 108.

Puigbo, P., et al. (2008), 'CAIcal: a combined set of tools to assess codon usage adaptation', *Biol Direct,* 3, 38.

Ranjan, A., et al. (2017), 'Signal recognition particle prevents N-terminal processing of bacterial membrane proteins', *Nat Commun,* 8, 15562.

Rudd, K. E. (2000), 'EcoGene: a genome sequence database for Escherichia coli K-12', *Nucleic Acids Res,* 28 (1), 60-4.

Sato, N. and Tajima, N. (2012), 'Statistics of N-terminal alignment as a guide for refining prokaryotic gene annotation', *Genomics,* 99 (3), 138-43.

Sberro, H., et al. (2019), 'Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes', *Cell,* 178 (5), 1245-59 e14.

Schrader, J. M., et al. (2014), 'The coding and noncoding architecture of the Caulobacter crescentus genome', *PLoS Genet,* 10 (7), e1004463.
Shine, J. and Dalgarno, L. (1975), 'Determinant of cistron specificity in bacterial ribosomes', *Nature,* 254 (5495), 34-8.
Smith, L. M., et al. (2013), 'Proteoform: a single term describing protein complexity', *Nat Methods,* 10 (3), 186-7.
Solbiati, J., et al. (1999), 'Processing of the N termini of nascent polypeptide chains requires deformylation prior to methionine removal', *J Mol Biol,* 290 (3), 607-14.
Staes, A., et al. (2011), 'Selecting protein N-terminal peptides by combined fractional diagonal chromatography', *Nat Protoc,* 6 (8), 1130-41.
Stekhoven, D. J., et al. (2014), 'Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism', *J Proteomics,* 99, 123-37.
Tatusova, T., et al. (2016), 'NCBI prokaryotic genome annotation pipeline', *Nucleic Acids Res,* 44 (14), 6614-24.
Thein, M., et al. (2010), 'Efficient subfractionation of gram-negative bacteria for proteomics studies', *J Proteome Res,* 9 (12), 6135-47.
Thomas, J. C., et al. (2006), 'A second isoform of the ferredoxin:NADP oxidoreductase generated by an in-frame initiation of translation', *Proc Natl Acad Sci U S A,* 103 (48), 18368-73.
Tobias, J. W., et al. (1991), 'The N-end rule in bacteria', *Science,* 254 (5036), 1374-7.
Tripp, H. J., et al. (2015), 'Toward a standard in structural genome annotation for prokaryotes', *Stand Genomic Sci,* 10, 45.
Tsirigos, K. D., et al. (2015), 'The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides', *Nucleic Acids Res,* 43 (W1), W401-7.
Van Damme, P., et al. (2014), 'N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men', *Mol Cell Proteomics,* 13 (5), 1245-61.
VanDrisse, C. M. and Escalante-Semerena, J. C. (2019), 'Protein Acetylation in Bacteria', *Annu Rev Microbiol*.
Varcamonti, M., et al. (2001), 'Proteins of the lactococcin A secretion system: lcnD encodes two in-frame proteins', *FEMS Microbiol Lett,* 204 (2), 259-63.
Verbruggen, S., et al. (2019), 'PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms', *Mol Cell Proteomics*.
Wang, H. and Matsumura, P. (1997), 'Phosphorylating and dephosphorylating protein complexes in bacterial chemotaxis', *J Bacteriol,* 179 (1), 287-9.
Weaver, J., et al. (2019), 'Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes', *MBio,* 10 (2).
Wilson, D. N. (2009), 'The A-Z of bacterial translation inhibitors', *Crit Rev Biochem Mol Biol,* 44 (6), 393-433.
Xue, Y. and Sherman, D. H. (2000), 'Alternative modular polyketide synthase expression controls macrolactone structure', *Nature,* 403 (6769), 571-5.
Yu, X. J., et al. (2011), 'Tandem translation generates a chaperone for the Salmonella type III secretion system protein SsaQ', *J Biol Chem,* 286 (41), 36098-107.
Yuan, P., et al. (2018), 'Comparative Membrane Proteomics Reveals a Nonannotated E. coli Heat Shock Protein', *Biochemistry,* 57 (1), 56-60.
Zhao, L., et al. (2011), 'A proteogenomic analysis of Shigella flexneri using 2D LC-MALDI TOF/TOF', *BMC Genomics,* 12, 528.
Zheng, J., et al. (2017), 'Proteogenomic Analysis and Discovery of Immune Antigens in Mycobacterium vaccae', *Mol Cell Proteomics,* 16 (9), 1578-90.

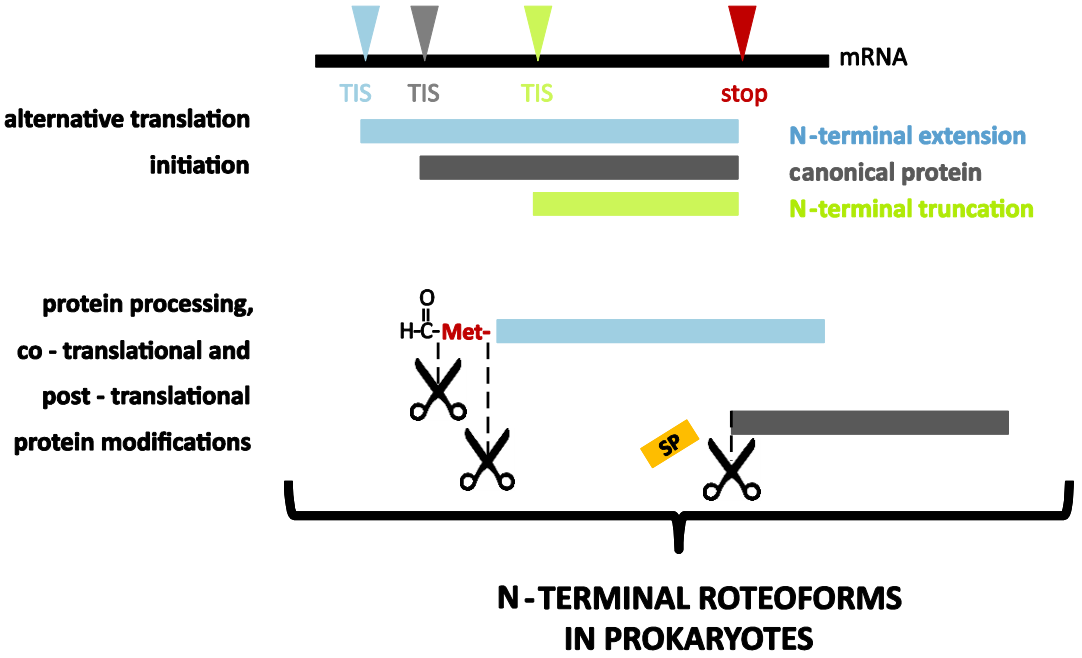

Figure 1| N-terminal proteoform expression of bacterial genes. The occurrence of alternative translation initiation (e.g. the use of in-frame translation initiation sites (TIS)) and N-terminal protein modifications (e.g. (partial)deformylation, (partial) methionine excision and signal peptide (SP) removal) lead to the expression of alternative N-terminal proteoform in bacteria.

Figure 2| Miss-annotation of bacterial translation initiation sites. Alignment of *E. coli* K-12, *Salmonella* 14028S and *Salmonella* SL1344 genomes using MAUVE (Darling et al. 2004) indicates expression of a conserved extended adhP proteoform (A) and an N-terminally truncated pyrD proteoform (B), which in case of the SL1344 strain were both predicted based on Ribo-seq expression evidence (REPARATION (Ndah et al. 2017))(A-B) and supported by N-terminal protein evidence in case of adhP (A) (Ndah et al. 2017). Height of MAUVE similarity profiles indicates average level of conservation of the genomic region, while the colour (green or purple) delineates separate blocks of detectable homology.

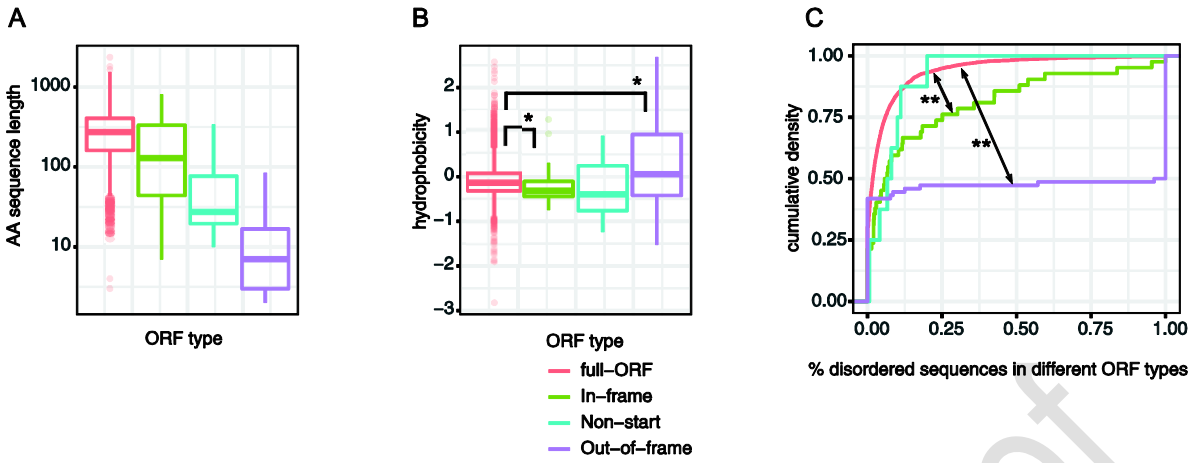

Figure 3| Properties of alternative coding sequences and their translation products predicted by retapamulin-assisted ribo-seq. Plots were generated using the 124 conserved TIS residing in annotated gene structures reported in (Meydan et al. 2019) using the *E. coli* BW25113 reference genome assembly (U00096.3). A) Different categories of internal TIS, namely in-frame (n=24), non-start (n=8) and out-of-frame (n=24) TIS belong to ORFs of decreased length compared to full-length annotated ORFs. B) Relationship between ORF type and Grand Average of Hydropathy (GRAVY) index (Kyte et al. 1982). C) Translation products of in-frame and out-of-frame ORFs contain a larger proportion of disordered sequence as compared to the canonical ORF translation product. * - Benjamini-Hochberg adjusted p value <0.01; ** adj. p-value <0.001; Wilcoxon-Mann-Whitney test for independent samples.

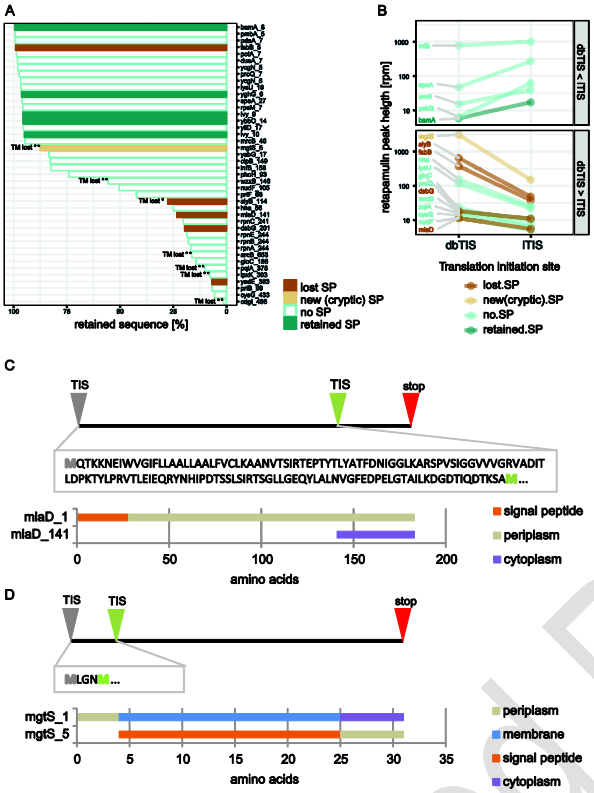

Figure 4| Differential Nt-proteoform expression revealed by inspecting publicly available Ribo-seq data though GWIPZ-vis. Expression of Nt-proteoforms identified using retapamulin-assisted Ribo-seq (Meydan et al. 2019) (see Figure 5) could be confirmed by inspecting tetracycline-assisted Ribo-seq data of (Nakahigashi et al. 2016). Relative expression of these proteoforms depend on *E. coli* growth conditions, according to data reported by Li et al. (Li et al. 2014).

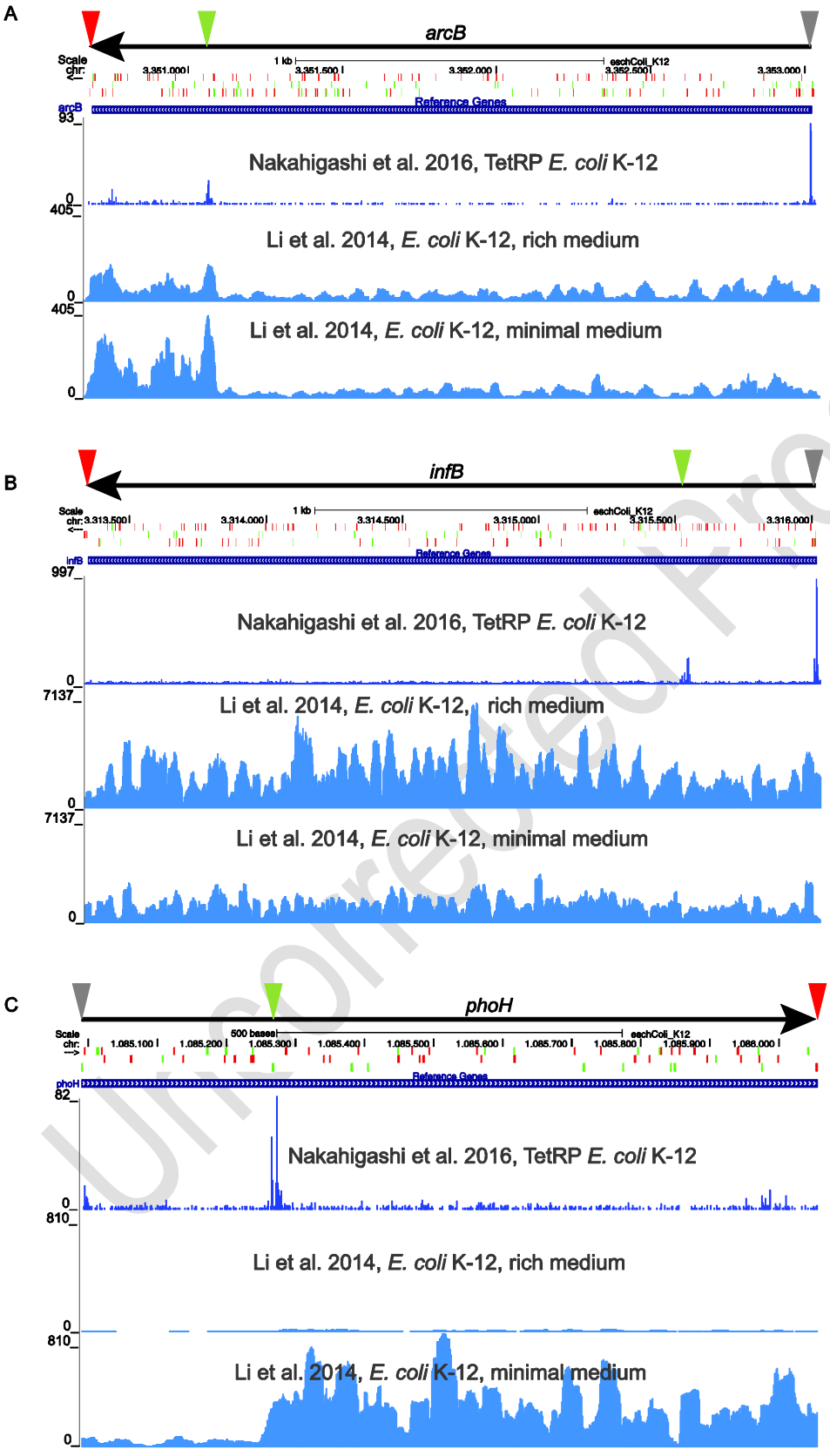

Figure 5| Comparison of proteoform pairs identified by retapamulin-assisted Ribo-seq in *E. coli* (N=42 pairs, Meydan et al. 2019). A) Every horizontal bar corresponds to a different truncated proteoform with their corresponding IDs as follows: 'gene name_internal start position (in amino

acids)'. Transmembrane regions were predicted and whenever one or more transmembrane domains were lost due to Nt-truncation, these are indicated with a „TM lost" label. * denotes TOPCONS (Tsirigos et al. 2015) prediction, while ** denotes both TOPCONS and Phobius prediction (Kall et al. 2004). SignalP prediction was colour-coded. „Lost SP" (dark brown) means that the canonical protein had a predicted SP in contrast to its Nt-truncated counterpart; „no.SP" (transparent) means neither the full-length nor Nt-truncated proteoform have a predicted SP; „retained SP" (green) means both CDS and Nt-proteoform harbour a predicted SP. In case of mgtS, the identified N-terminally truncated proteoform lost 4 N-terminal amino acids preceding an annotated transmembrane domain thereby revealing a "new (cryptic) SP" (light brown). B) The relationship between retapamulin peak height (in reads per million (rpm)) and TIS identity (internal TIS (iTIS); database-annotated TIS (dbTIS)). Illustrative examples of a mlaD proteoform losing a signal peptide (C) and a mgtS proteoform gaining a signal peptide (D) show the TOPCONS predicted topology of full-length and truncated proteoforms.